Prediction of 2024 US election ...*

Colin Sihan Yang Lexun Yu Siddharth Gowda

November 4, 2024

We forecast the winner of the 2024 US presidential election using "poll-of-polls" by building a linear model.

1 Introduction

Election result forecasting has become an essential tool for analysts in political science and the public to predict the outcome of democratic process, such as the presidential election in the United States. Traditionally, individual polls have been used as a snapshot of voter sentiment, but they only reflect temporary changes in the performance of contestants, instead of a precise estimation of the election result. As discussed by Pasek (2015) and Blumenthal (2014), the aggregation of multiple polls, or "poll-of-polls," has become a popular technique to reduce individual survey errors and provide more accurate election forecasts. However, the traditional poll aggregation does not reflect dynamics of an election, especially with real-time changes and the introduction of new data. This creates a gap for a more adaptable model to predict the election result based on both polling data and additional variables, such as historical data and economic indicators.

This paper fills the gap by building a hybrid election forecasting model following the strategies mentioned by Pasek (2015). As Pasek (2015) described in their article, aggregation involves determining which surveys are worth including, as well as selecting, combining and averaging results from multiple polls to reduce individual biases and errors. Prediction modeling adds other data to the model that predicts election outcomes based on current dynamics. Hybrid models like the Bayesian approach incorporates prior beliefs based on historical data or expert knowledge and new evidence like economic updates to dynamically adjust the forecast as the campaign progresses.

In this paper, we aim to predict the 2024 us election result with the hybrid election forcasting model. We incorporate aggregation by filtering the polls on FiveThirtyEight (2024) by

*Code and data are available at: https://github.com/yulexun/uselection.

numeric grade that indicates pollster's reliability, prediction that incorporates social and economic indicators including unemployment rates and abortion rates, and hybrid approaches that leverages Bayesian techniques which combines historical data such as the 2016 election data, allowing for a dynamic prediction of the U.S. presidential election.

The estimand for this research paper is the predicted support percentages for Kamala Harris and Donald Trump. The prediction is based on quantifying various polling factors, including sample size, poll scores, and transparency scores, which are used as predictors.

The results of this model indicate a more stable and accurate forecast compared to traditional aggregation methods alone, [update this …]

The remainder of this paper is structured as follows: [update this …]

Appendix

2 Additional data details

3 Model details

3.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

3.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

4 FiveThirtyEight Licenses

5 Trump Voter Prediction Model

The multiple linear regression model (MLR) for Donald Trump will use the same variables, formula, and Bayesian approach as the one for Harris. Likewise, the Trump dataset is also split into training and testing data. The model outputs are below.

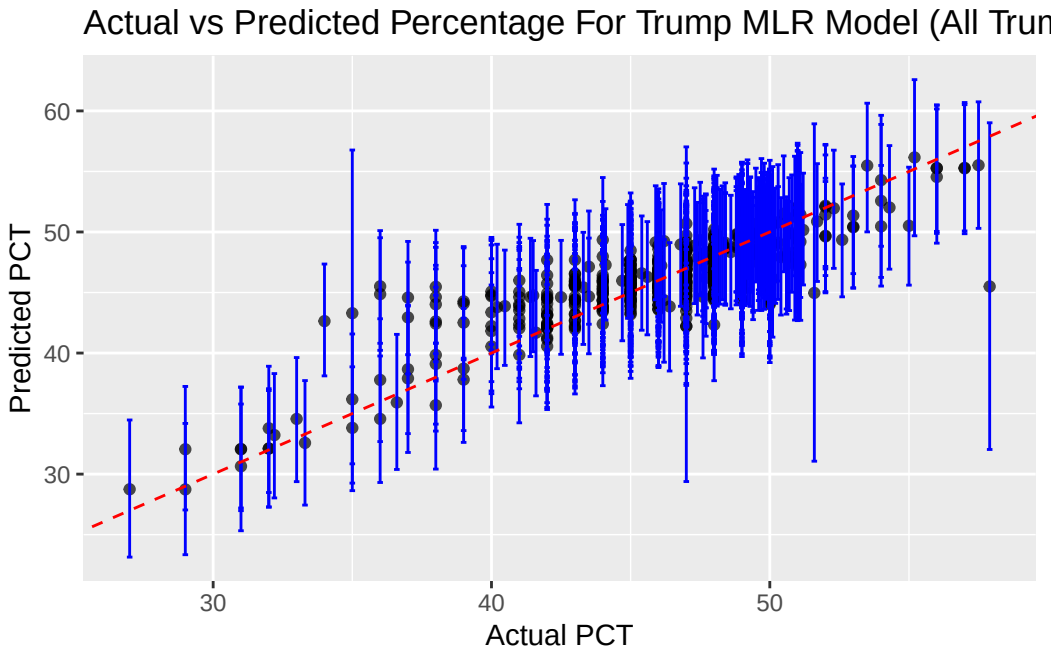

Figure 1: MLR Trump Model Accounts For A Large Amount of Variability in Voter Percentage

From Figure 1 it is clear that the model accounts for large amount of the variance in Trump's voter percentage as the data points appear close to the prediction (red line). Furthermore, the distance between the prediction and the actual values do not appear to follow a pattern, suggesting that the error is due to randomness and not model bias.

Based on Figure 2, the test data predictions are also close to the actual values. This suggests that model can generalization to outside data. Similarly, the distance between the prediction and the actual values do not appear to follow a pattern, suggesting that the error is due to randomness and not model bias.

Based on Figure 3, the model expects Trump to win slight less than 45% of the popular vote and the 95% confidence interval ranges from around 42.5% to 47.2%. This confidence interval is slightly larger than the model for Harris, implying that the Trump polling data might be less reliable.

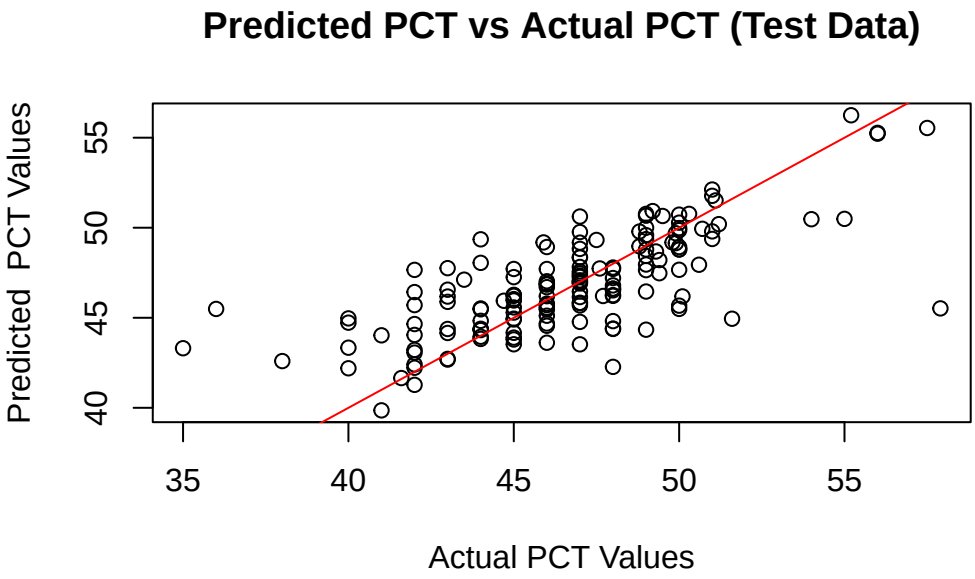

Figure 2: MLR Trump Model does not Appear to Overfit

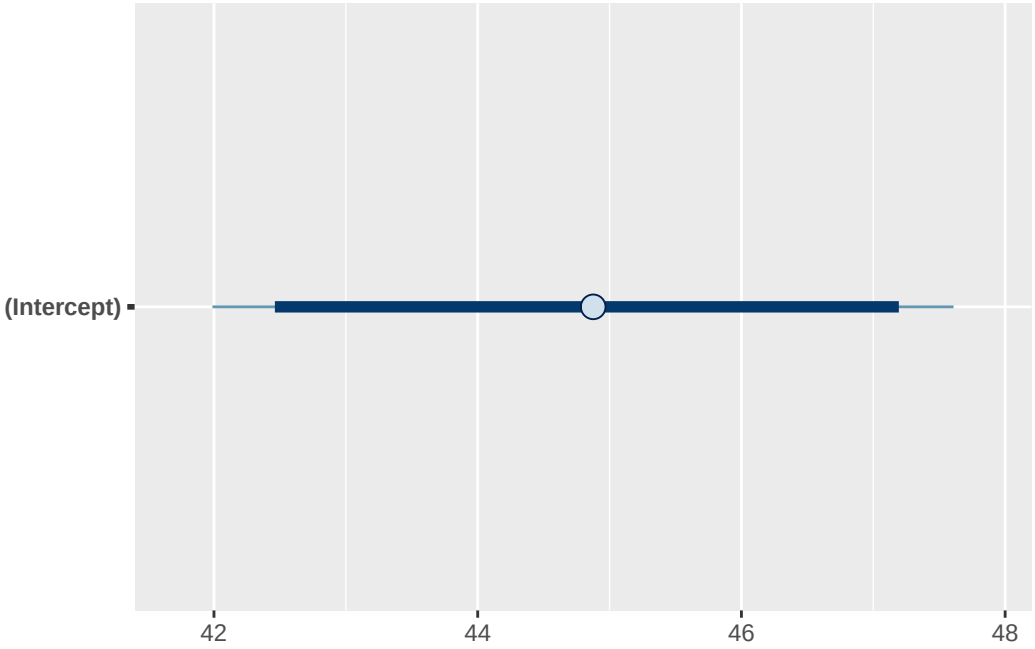

Figure 3: Donald Trump Expected to Recieve Approximatley 45% of the Vote

6 Election Prediction

Our prediction process consists of two primary components. First, we develop models for both Trump and Harris based on the variables outlined in Section **?@sec-PLACEHOLDER**. This involves partitioning the dataset into training and testing subsets. Next, we further divide the testing dataset into swing states and other competitive races. We then input this test data into the respective models to generate predictions. By averaging these predictions, we can calculate the expected voter percentage for each candidate in each state. The candidate with the higher percentage is deemed the winner for that state.

We generated predictions for the following states: Arizona, Nevada, Georgia, Pennsylvania, Michigan, Minnesota, Wisconsin, Florida, Texas, Maine CD-2, Nebraska CD-2, New Hampshire, Ohio, Virginia, North Carolina, and Iowa. Winners for other states were determined based on historical trends and predictions from sources like (**cnn?**). Most states without predictions are strongly Republican or Democratic, so their absence is not expected to significantly impact prediction validity.

Table 1: Kamala Harris Wins Most of the Swing States

State	Harris Predicted Percentage	Trump Predicted Percentage	State Winner
Arizona	46.60627	49.20840	Trump
Florida	42.86374	50.74617	Trump
Georgia	47.21810	48.86554	Trump
Iowa	48.56302	43.52854	Harris
Maine CD-2	47.30270	49.28971	Trump
Michigan	47.59810	46.99337	Harris
Minnesota	48.62212	43.66140	Harris
Nebraska CD-2	49.95245	42.15686	Harris
Nevada	49.35471	47.11387	Harris
New Hampshire	50.78576	42.63839	Harris
North Carolina	48.68842	47.73255	Harris
Ohio	43.96915	50.97722	Trump
Pennsylvania	48.20624	47.17918	Harris
Texas	44.97007	50.26199	Trump
Virginia	49.20564	43.19202	Harris
Wisconsin	48.44502	46.45388	Harris

According to Figure **?@fig-currentstateofelection**, Kamala Harris is predicted to be the 47th President of the United States. There are also a few states with predictions not visible in the map, we will describe those predicts them below

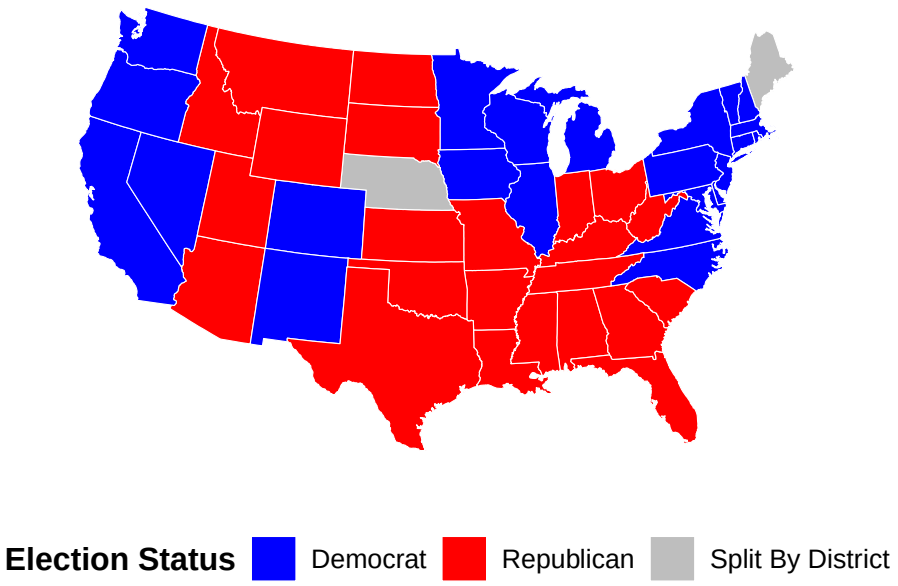

Figure 4: Kamala Harris is Predicited to be the 47th President of the United States

In Maine, Harris is projected to win the state's overall delegates and District 1, while District 2 is expected to go to Trump. In Nebraska, Trump is expected to win the state's delegates along with Districts 1 and 3, while Harris is predicted to win District 2. Additionally, Trump is projected to win Alaska, and Harris is expected to win Hawaii.

Overall, the predictions indicate that Harris will receive 298 delegates, while Trump will receive 240 delegates.

7 Sources of Bias In the Polls

Based on the figures in section **?@sec-data**, the team identified three sources of bias beyond sample size and state: methodology, the timing of polls relative to the election date, and poll scores. From the analysis in Figure **?@fig-methodologyboxplot**, we found that polls employing a limited number of methods to reach and communicate with voters exhibited greater variability. This finding is plausible, as different demographic groups often prefer distinct communication methods; for example, older voters may favor phone surveys, while younger voters might prefer online surveys. Additionally, utilizing multiple communication channels provides pollsters with more opportunities to contact potential voters, thereby reducing non-response bias. Consequently, it is essential to incorporate methodology as a variable in the model, and we recommend that pollsters adopt diverse communication methods whenever possible.

From Figure **?@fig-daysfromelection-pct**, the team observed a decline in third-party candidate support as the election approached. Early in the election cycle, voters may entertain the idea of a viable third-party candidate, but as the election nears, interest in third-party options diminishes. This trend may be worsened by decreased participation in polls as the election date approaches. As a result, the people who do participate in those earlier polls are likely to be more informed and passionate, which are qualities that may be associated with support for third-party candidates. Additionally, the withdrawal of a prominent third-party candidate, Robert F. Kennedy Jr., on August 23, 2024 (74 days from the election), could influence this dynamic. However, we find the earlier reasons more compelling, as the data in Figure **?@fig-daysfromelection-pct** do not indicate a significant drop in third-party support at that time, but rather a gradual decline.

As illustrated in Figure **?@fig-pollscore-pct**, more reliable polls tend to show a bias favoring Trump over Harris—not necessarily indicating a guaranteed victory for Trump, but rather suggesting he may outperform expectations in many polls. It is important to note that the 538 poll score system is partly based on historical accuracy. Given Trump's overperformance in the polls during the 2016 and 2020 elections, this scoring may be overcorrecting for past inaccuracies. This potential undervaluation of Trump's support could be strategically beneficial for the Harris campaign, providing a warning to avoid repeating previous electoral mistakes.

7.1 Limitations of the Model

While this model presents a more comprehensive approach, there are limitations. First, Any systematic biases inherent in those polls can carry over into the model's predictions because we are using existing poll data. For instance, if certain demographic groups are consistently underrepresented, our model may not fully capture their impact on election outcomes.

Second, the model only reflect current voting dynamics as it is based on historical data. This model does not account for short-term variability such as Biden's exit. This limitation is

important in the final weeks before an election when small shifts in polling data can produce exaggerated effects.

On top of that, the results can be sensitive to the choice of priors. We choose normal(0, 2.5) as we believe it could stabilize the estimates by being informative enough to guide the posterior distribution without overpowering the data. However, if the chosen prior does not align well with the true underlying distribution, it may lead to biased results or overly wide credible intervals.

Also, as it is a model based on a dataset, it is challenging to capture all the real-life features and dynamics. The model might not include all relevant predictors or interactions due to limitations in data availability or complexity, which can lead to incomplete representations of the factors influencing voting behavior. Additionally, without strong regularization techniques, the model may become prone to overfitting, particularly when using complex hierarchical structures or including numerous predictors. This overfitting can reduce the model's generalizability to new or unseen data.

Moreover, errors and offsets inherent in polling data, such as response bias, nonresponse adjustments, and sampling variability, can propagate into the model's results. These aspects introduce an additional layer of uncertainty that can affect the model's reliability and predictive performance. While Bayesian methods provide a robust framework to incorporate uncertainty, the final outputs must be interpreted cautiously, acknowledging these underlying limitations.

8 Limitations of the Prediction

In section **?@sec-prediction**, the team predicted that Kamala Harris would secure 298 delegates while Donald Trump would obtain 240 delegates. However, this prediction has several limitations. Firstly, the model used to generate these predictions inherits the constraints outlined in section **?@sec-model-limitations**. Additionally, the predictions are based on averaging results from polling data for each state, which can lead to propagated prediction errors.

Each state has a limited number of polls available; while the overall analysis considered a substantial number of polls, individual states often had fewer than 20 polls, with even the largest states contributing only up to 27. Moreover, as noted in Section **?@sec-prediction**, the analysis focused solely on closely contested races, specifically swing states. Consequently, many states were assigned to a candidate based on strong historical trends rather than direct predictions. Although this approach is reasonable, it may overlook emerging trends that could potentially shift a state's allegiance from Republican to Democrat or vice versa.

References

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators." *PS: Political Science and Politics* 47 (2): 297–300. http://www.jstor.org/stable/43284537.

FiveThirtyEight. 2024. "Our Data." *FiveThirtyEight*. https://data.fivethirtyeight.com.

Pasek, Josh. 2015. "THE POLLS–REVIEW: PREDICTING ELECTIONS: CONSIDERING TOOLS TO POOL THE POLLS." *The Public Opinion Quarterly* 79 (2): 594–619. http://www.jstor.org/stable/24546379.